Ensemble Density-Functional Theory for *Ab Initio* Molecular Dynamics of Metals and Finite-Temperature Insulators

Nicola Marzari,[1,2] David Vanderbilt,[1] and M. C. Payne[2]
[1]*Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08855-0849*
[2]*Cavendish Laboratory (TCM), University of Cambridge, Madingley Road, Cambridge CB3 0HE, United Kingdom*
(Received 7 March 1997)

A new method is presented for performing first-principles molecular-dynamics simulations of systems with variable occupancies. We adopt a matrix representation for the one-particle statistical operator $\hat{\Gamma}$ to introduce a "projected" free energy functional G that depends on the Kohn-Sham orbitals only and that is invariant under their unitary transformations. The Liouville equation $[\hat{\Gamma}, \hat{H}] = 0$ is always satisfied, guaranteeing a very efficient and robust variational minimization algorithm, that can also be extended to nonconventional entropic formulations. [S0031-9007(97)03810-6]

PACS numbers: 71.15.Pd, 31.50.+w, 65.50.+m, 68.35.Ja

In recent years, the range of problems that can be studied with quantitative accuracy using the methods of computational solid state physics has expanded dramatically. It is now possible to calculate many materials properties with a precision that is often comparable to that of experiments. This degree of confidence is based on the fundamental quantum-mechanical treatment offered by density-functional theory (DFT) [1], coupled with the availability of increasingly powerful computers and with the development of algorithms tuned towards optimal performance [2,3].

The application of these methods and techniques to metallic systems has nonetheless encountered several difficulties that have made progress slower than for the case of semiconductors and insulators. The discontinuous variation of the orbital occupancies across the Brillouin zone (BZ) makes the occupation numbers rather ill-conditioned variables, and the self-consistent solution of the screening problem can suffer from several instabilities. The absence of a gap in the energy spectrum and the requirement of an exact diagonalization for the Hamiltonian matrix everywhere in the BZ (in order to assign the occupation numbers) introduce "slow frequencies" in the evolution of the orbitals towards the ground state and preclude the straightforward extension to metals of algorithms which performed well for insulators. Smearing the Fermi surface with a finite electronic temperature [4,5] allows for an improved BZ sampling, but only partially alleviates the problems alluded to above.

In this Letter, we introduce a new approach which solves many of these problems in a natural way, and which provides a general and efficient framework for obtaining the ground state of a Kohn-Sham Hamiltonian at a finite electronic temperature. The context is the Mermin formulation for the Fermi-Dirac statistics [5–7], but the method also applies when generalized entropic functionals are introduced [8], as is often the case for metallic systems. The language of ensemble DFT [9] is used, and a variational algorithm for the minimization to the ground state is developed and implemented. Dramatic improvements are obtained in the convergence of the energies and especially of the Hellmann-Feynman forces.

Within ensemble DFT, the Helmholtz free energy functional at a temperature T and for an N-representable charge density $n(\mathbf{r})$ in an external potential V_{ext} is $A_V[n(\mathbf{r})] = F_T[n(\mathbf{r})] + \int V_{\text{ext}}(\mathbf{r}) n(\mathbf{r})\, d\mathbf{r}$, where F_T is the finite-temperature Mermin-Hohenberg-Kohn functional [6]. The charge density $n_0(\mathbf{r})$ that minimizes A_V is the ground-state charge density, and $A_V[n_0(\mathbf{r})]$ is the free energy of the electronic system. A Kohn-Sham mapping onto noninteracting electrons leads to a decomposition of the functional F_T into noninteracting kinetic-energy, electrostatic, and entropic contributions, plus the exchange-correlation (XC) functional, for which we take the local density approximation [1].

A key assumption is made by adopting a matrix representation f_{ij}, in the basis of the orbitals, for the one-particle effective statistical operator $\hat{\Gamma}$, so that

$$n(\mathbf{r}) = \sum_{ij} f_{ji}\, \psi_i^*(\mathbf{r}) \psi_j(\mathbf{r})\,. \tag{1}$$

Here the $\{\psi_i\}$ are orthonormal single-particle Kohn-Sham orbitals, the sum extends in principle over all the states, and f_{ij} is constrained to have $\operatorname{tr} \mathbf{f} = N$ and eigenvalues bounded to $[0, 1]$. The functional A to be minimized is then

$$A[T; \{\psi_i\}, \{f_{ij}\}] = \sum_{ij} f_{ji} \langle \psi_i | \hat{T} + \hat{V}_{\text{ext}} | \psi_j \rangle + E_{\text{H}XC}[n] - TS[\{f_{ij}\}]; \tag{2}$$

the Hartree and XC terms, which depend only on the charge density, have been grouped together. The Fermi-Dirac entropic term is a function of the eigenvalues of $\mathbf{f}$: $S[\{f_{ij}\}] = \operatorname{tr} s(\mathbf{f})$, where s is $f \ln f + (1 - f) \ln(1 - f)$. Typically, V_{ext} is generated by an array of nonlocal pseudopotentials. The free energy functional A is in the form of traces of operators, and so it is covariant under a change of representation (i.e., for a unitary transformation $\mathbf{U}$ of the orbitals $\{\psi_i\}$); this can be verified by letting $\mathbf{f} \rightarrow \mathbf{f}' = \mathbf{U}\mathbf{f}\mathbf{U}^\dagger$ and $|\psi_j\rangle \rightarrow |\psi_j\rangle' = \sum_m U^*_{jm}\psi_m$. The covariance of A allows for the definition of a new *projected functional* G that depends only on the orbitals $\{\psi_i\}$:

$$G[T;\{\psi_i\}] \coloneqq \min_{\{f_{ij}\}} A[T;\{\psi_i\},\{f_{ij}\}]\,. \tag{3}$$

G is *invariant* under any unitary transformation of the $\{\psi_i\}$: the transformed orbitals cannot lead to a different value for G, by virtue of the covariance of A.

The projected functional G represents a much better conditioned choice than the original free energy A for the evolution of the orbitals towards the ground state. The reasons are several, albeit related. (i) The functional G no longer depends on the occupancies of the orbitals or on their unitary transformations ("rotations") in the occupied subspace. These are ill-conditioned, nonlocal degrees of freedom, with the added nonlinear constraint of charge normalization. (ii) The f_{ij} have become dependent variables, implicitly defined by the minimization in Eq. (3), and this dependency does not enter into the calculation of the functional derivatives $\delta G/\delta\psi_i^*$, since the contributions $(\partial G/\partial f_{kl})(\partial f_{kl}/\partial\psi_i^*)$ are zero because of the minimum condition. (iii) The occupancies of the orbitals and their internal rotations are now consistently considered as part of the same problem (finding the ground state $\hat{\Gamma}$), and not as two independent problems. (iv) The expensive and inefficient evolution for the orbital rotations is now shifted to the matrix $\mathbf{f}$; this implies that the associated slow frequencies [10] have now been removed (compressed to zero) by the minimization condition. Subspace alignment [11,12] between subsequent orbital updates is automatically enforced.

This formulation naturally decouples the evolution of the orbitals $\{\psi_i\}$ from that of the f_{ij}: the orbitals get updated as an outer loop minimizes G, and after each update an inner loop on the f_{ij} minimizes *A at fixed orbitals* $\{\psi_i\}$. The minimization of G is then freed from all constraints but the orthonormality of the $\{\psi_i\}$. By contrast, conventional methods adopt a diagonal representation for $\hat{\Gamma}$ (the occupation numbers f_i), and add orbitals rotations and updates of the f_i [5,10,13], or diagonalizations of an updated Hamiltonian [12], in order to reach the ground state. Such approaches have several drawbacks. The rotations (or diagonalizations) are not guaranteed to lower the free energy, unless additional constraints are imposed [5], and do not take into account the rotation dependence of $n(\mathbf{r})$ (directly, or indirectly via the changes it induces in the self-consistent Hamiltonian). The dynamics of rotation introduces poorly conditioned slow frequencies in the evolution of the orbitals [10]. Also, the $\{\psi_i\}$ are made inequivalent by association with different f_i, thus precluding the subspace alignment [11,12] between successive orbital updates and interfering with a meaningful minimization algorithm. The behavior of A along the search lines becomes increasingly complicated in larger systems, leading to inaccurate minima.

Returning to our approach, and beginning with the inner-loop problem, let us define

$$h_{ij} = \langle\psi_i|\hat{T} + \hat{V}_{\rm ext}|\psi_j\rangle, \qquad V_{ij}^{[n]} = \langle\psi_i|\hat{V}_{\rm HXC}^{[n]}|\psi_j\rangle$$

for the matrix elements of the Hamiltonian (the superscript $[n]$ is a reminder of the dependence on charge density); the minimum condition that defines G implies

$$\begin{aligned}\frac{\delta A}{\delta f_{ji}} &= h_{ij} + \frac{\delta E_{\rm HXC}}{\delta f_{ji}} - T\,\frac{\delta S}{\delta f_{ji}} - \mu\delta_{ij} \\ &= h_{ij} + \int d\mathbf{r}\,\frac{\delta E_{\rm HXC}}{\delta n(\mathbf{r})}\,\frac{\delta n(\mathbf{r})}{\delta f_{ji}} - T[s'(\mathbf{f})]_{ij} - \mu\delta_{ij} \\ &= h_{ij} + V_{ij}^{[n]} - T[s'(\mathbf{f})]_{ij} - \mu\delta_{ij} = 0\,. \end{aligned} \tag{4}$$

The constraint of charge conservation $\mathbf{tr}\,\mathbf{f} = N$ is enforced via the Lagrange multiplier μ, and the notation $[s'(\mathbf{f})]_{ij}$ is used in place of d tr $s(\mathbf{f})/df_{ji}$. The stationarity condition in (4) implies that f_{ij} and the Hamiltonian $h_{ij} + V_{ij}^{[n]}$ are diagonalized by the same unitary rotation, at fixed orbitals, and thus represent "commuting" operators; the non-self-consistent Liouville equation $[\hat{\Gamma},\hat{H}] = 0$ is satisfied. The relation (4) does not mean that $h_{ij} + V_{ij}^{[n]}$ and f_{ij} are diagonal, but that there is a common transformation that diagonalizes both; the formalism is not linked to a preferred diagonal representation.

The inner loop for the update of the occupation matrix f_{ij} is carried out at fixed orbitals, and so it does not require the calculation of new matrix elements for the kinetic energy operator or the nonlocal pseudopotential, and there are no orthogonalizations involved. The Fourier transforms of the $\{\psi_i\}$ can also be eliminated by storing their real-space representation. Note that if the problem were not self-consistent, the solution for the equilibrium f_{ij} would be found by straightforwardly diagonalizing the Hamiltonian matrix, calculating from its eigenvalues the thermal distribution of the occupation numbers, and rotating these back into the current orbital representation. This is not the solution, but we use it as a search direction for a direct line minimization in the multidimensional space of the f_{ij}. The procedure is organized as follows. The matrix h_{ij} is determined once for all before entering the inner loop. The updated charge density (on the mth iteration in the inner loop) is

$$n^{(m)}(\mathbf{r}) = \sum_{ij} f_{ji}^{(m)}\psi_i^*(\mathbf{r})\psi_j(\mathbf{r})\,. \tag{5}$$

$E_{\rm HXC}^{(m)}$ and $V_{\rm HXC}^{(m)}(\mathbf{r})$ are then calculated, and the matrix representation $V_{ij}^{(m)}$ is constructed. The entropy $S^{(m)}$ is also computed, following a diagonalization of $\mathbf{f}$,

$$f_{ij}^{(m)} = \sum_l Y_{il}^{(m)\dagger} f_l^{(m)} Y_{lj}^{(m)}. \tag{6}$$

In a plane-wave approach the charge density (5) is calculated more efficiently in this representation in which the f_{ij} are diagonal, since a temporary rotation of the orbitals can then be performed on their more compact reciprocal-space representation. The Hamiltonian matrix is updated with the new local terms, and diagonalized:

$$H_{ij}^{(m)} = h_{ij} + V_{ij}^{(m)} = \sum_l Z_{il}^{(m)\dagger} \epsilon_l^{(m)} Z_{lj}^{(m)}. \tag{7}$$

The non-self-consistent minimum for $\mathbf{f}$ would be

$$\tilde{f}_{ij}^{(m)} = \sum_l Z_{il}^{(m)\dagger} f_T(\epsilon_l^{(m)} - \mu) Z_{lj}^{(m)}, \tag{8}$$

where f_T is the (Fermi-Dirac) thermal distribution. We choose this as our search direction in the f_{ij} space, and a full line minimization is performed along the multidimensional segment $\mathbf{f}_\beta^{(m+1)} = \mathbf{f}^{(m)} + \beta \Delta \mathbf{f}^{(m)}$, where $\Delta \mathbf{f}^{(m)} = \tilde{\mathbf{f}}^{(m)} - \mathbf{f}^{(m)}$. Note that β parametrizes an *unconstrained* search, since tr $\mathbf{f}_\beta = N$ at the end points and thus, by linearity, at all β. Since the search direction is determined by the eigenvalues and eigenvectors of $\hat{H}$, and not from the occupation numbers, this current formalism can also be applied when generalized entropic functionals or non-monotonic thermal distributions are introduced [8].

The minimization proceeds by calculating the free energy and its derivative along the search line at the two end points $\beta = 0$ and $\beta = 1$, *taking into account the self-consistent variations in the charge density* and thus in the Kohn-Sham Hamiltonian. The line derivative is A' is $\sum_{ij} \Delta f_{ji}^{(m)} (\delta A / \delta f_{ji})$, where

$$\left.\frac{\delta A}{\delta f_{ji}}\right|_{\beta=0} = \left[h_{ij} + V_{ij}^{(m)} - T \sum_l Y_{il}^{(m)\dagger} s'(f_l^{(m)}) Y_{lj}^{(m)} \right];$$

$A'(\beta = 0)$ is always smaller than 0, and so the iterative update of $\mathbf{f}$ takes place in a strictly variational fashion. We then calculate the charge density $\tilde{n}^{(m)}(\mathbf{r})$, the matrix elements $\tilde{V}_{ij}^{(m)}$, and the free energy $\tilde{A}$ corresponding to $\beta = 1$, together with the line derivative $A'(\beta = 1)$ via

$$\left.\frac{\delta A}{\delta f_{ji}}\right|_{\beta=1} = \left[h_{ij} + \tilde{V}_{ij}^{(m)} - T \sum_l Z_{il}^{(m)\dagger} s'(\tilde{f}_l^{(m)}) Z_{lj}^{(m)} \right].$$

Since the $\hat{T}$ and $\hat{V}_{\text{ext}}$ contributions are exactly linear along the search direction, and the Hartree energy is quadratic, while the remaining XC and entropic terms are very well behaved, a cubic or a parabolic interpolation locates the optimal β with very good accuracy. More importantly, the choice of a direct minimization for $\mathbf{f}$ along a linear search implies that *level-crossing instabilities are completely eliminated*, even in the limit of zero temperature. In practice, we find that two iterations in the inner loop are an optimal choice [14], since we ultimately need self-consistency also with the $\{\psi_i\}$.

In the outer loop, G can be minimized efficiently with a direct all-bands conjugate gradient method. However, due to the broader spectrum that occupancies approaching zero introduce, the convergence is slower than for systems with a gap. Thus, we resort to a preconditioning strategy: We choose a set of scaled variables in which the functional has a more compressed spectrum, and the search directions are calculated in this new metric. In the diagonal representation, the total energy around the minimum is a quadratic form $\sum_{i,n} f_i \epsilon_i c_{i,n}^2$ ($c_{i,n}$ is the expansion coefficient of ψ_i in the nth element of the basis set); if we take the gradients according to the scaled variables $\tilde{c}_i = \sqrt{f_i}\, c_{i,n}$, the (unconstrained) preconditioned gradients for the original variables become [15]

$$-\frac{\delta G}{\delta \psi_i^*} \to -\frac{1}{f_i} \frac{\delta G}{\delta \psi_i^*} = -\hat{H} \psi_i . \tag{9}$$

With some degree of overcorrection [16], these can also be used to construct conjugate directions. A generalization to our case is obtained by calculating the gradients of G, passing them into the diagonal representation where they can be preconditioned as in (9), and transforming them back in the initial representation. The gradients $\mathbf{g}_i$ are

$$\mathbf{g}_i = -\frac{\delta G}{\delta \psi_i^*} = -\sum_j f_{ji} \hat{H} |\psi_j\rangle; \qquad \mathbf{g}_i' = -f_{ii}' \hat{H} |\psi_i'\rangle , \tag{10}$$

where the primed term refers to the diagonal representation ($\mathbf{f}' = f_{ii}' \delta_{ij} = \mathbf{U}\mathbf{f}\mathbf{U}^\dagger$). The preconditioned gradients $\mathbf{G}_i'$ and $\mathbf{G}_i$ are thus

$$\mathbf{G}_i' = -\hat{H} |\psi_i'\rangle = -\hat{H} \left(\sum_m U_{im}^* |\psi_m\rangle \right), \tag{11}$$

$$\mathbf{G}_i = \sum_n U_{in}^{*\dagger} \mathbf{G}_n' = -\hat{H} |\psi_i\rangle . \tag{12}$$

Such preconditioned $\mathbf{G}_i$ greatly improve the convergence rate, and are much cheaper to compute than the $\mathbf{g}_i$ in Eq. (10). A standard kinetic-energy preconditioning [2] should also be used in plane-wave calculations. In summary, on each iteration (i) each preconditioned gradient $-\hat{H}|\psi_i\rangle$ is calculated, conjugated with the previous search direction, and projected out of the subspace spanned by the orbitals to impose orthonormality (to first order) along the search; (ii) the first derivative of the free energy along the multidimensional (all bands, all plane waves, and all $\mathbf{k}$ points) line is calculated, and a trial step along the search line is taken; and (iii) after reorthogonalizing the orbitals, the new free energy provides the third constraint to identify the minimum along the search line.

The complete algorithm provides a remarkably robust and efficient convergence, and, at variance with iterative methods [3], it is formulated in a strictly variational fashion, where the free energy is bound to converge towards the ground state after every iteration. As a paradigmatic case we present here results for a unit cell that is 32 Å long, and contains a 15-layer 1×1 Al(110) slab. We use the single $\mathbf{k}$ point $\frac{2\pi}{a_0}(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, a fictitious

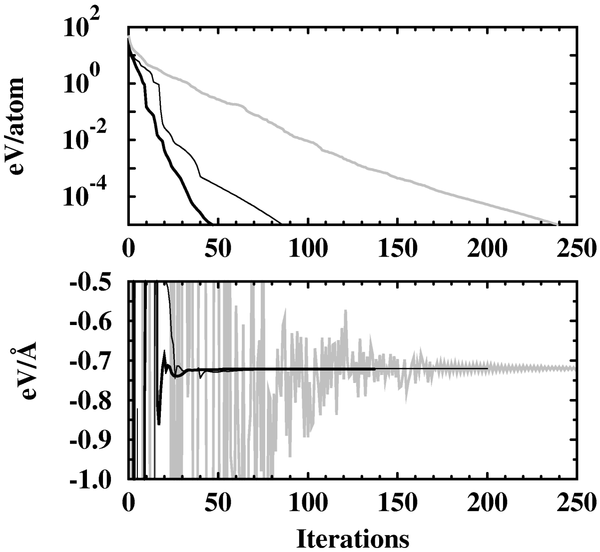

FIG. 1. Convergence of the total free energy (upper panel) and of the force on the surface atom (lower panel) in the 15-layer Al(110) slab. Grey line, ABV; thin (thick) solid line, ensemble DFT with two (four) iterations in the inner loop.

Gaussian temperature [13] of 4 eV, 100 eV plane-wave cutoff, and 64 orbitals. The large value of the temperature assures that the coarse sampling is sufficient; similar results are obtained with smaller and more physical temperatures. Figure 1 monitors the convergence of the total free energy and of the Hellmann-Feynman forces as a function of the number of iterations on the $\{\psi_i\}$; we compare an optimal all-bands variational implementation (ABV) [13] with the scheme that we have described (ensemble DFT) [17]. The improved convergence for the energies, but especially for the Hellmann-Feynman forces, is clearly apparent. It should be stressed that, at variance with the ABV case, the free energy along the line searches in both the outer and inner loops is accurately parabolic; interpolated minima are usually fractions of a percent off their true values. The improved convergence of the ionic forces leads to a much tighter conservation of the constant of motion in molecular dynamics simulations. We show in Fig. 2 the results of a run for our Al(110) slab. The time step is 2 fs; the ions are moved after a fixed tolerance in the convergence of the free energy is reached (identical results are obtained if a fixed number of iterations is used). The systematic drift of the constant of motion stabilizes after $\sim$0.3 ps of thermalization to -0.6 (eV/cell)/ps for the ABV case, and to -0.0008 (eV/cell)/ps for ensemble DFT. Such stability opens the way to inexpensive molecular dynamics simulation of large metallic systems even on common workstations [14].

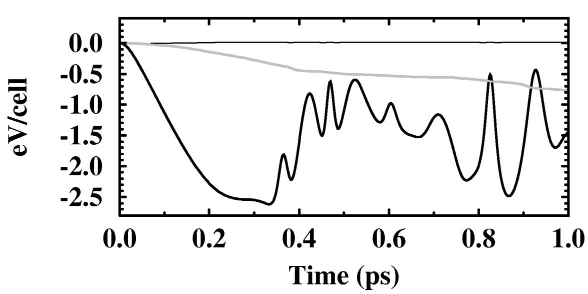

FIG. 2. Conservation of the constant of motion in a molecular-dynamics run for the 15-layer Al(110) slab. Bottom curve, electronic free energy; top two curves, same plus ionic kinetic energy (grey line, ABV; solid thin line, ensemble DFT with two iterations in the inner loop).

N. M. thanks A. De Vita and J. White for many helpful discussions. We acknowledge support from Grant No. ERBCHBICT920192 (N. M.) and NSF Grant No. DMR-91-15342 (D. V.), and access to the Hitachi S3600, Hitachi Europe, Maidenhead, UK.

[1] See, e.g., R. O. Jones and O. Gunnarsson, Rev. Mod. Phys. **61**, 689 (1989).

[2] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985); M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992); D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).

[3] A. Alavi, J. Kohanoff, M. Parrinello, and D. Frenkel, Phys. Rev. Lett. **73**, 2599 (1994); G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11 169 (1996).

[4] C.-L. Fu and K.-M. Ho, Phys. Rev. B **28**, 5480 (1983).

[5] M. J. Gillan, J. Phys. Condens. Matter **1**, 689 (1989).

[6] N. D. Mermin, Phys. Rev. **137**, A1441 (1965).

[7] R. M. Wentzcovitch, J. L. Martins, and P. B. Allen, Phys. Rev. B **45**, 11 372 (1992).

[8] A. De Vita, Ph.D. Thesis, University of Keele, 1992; N. Marzari *et al.* (to be published).

[9] See, e.g., R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).

[10] M. P. Grumbach, D. Hohl, R. M. Martin, and R. Car, J. Phys. Condens. Matter **6**, 1999 (1994).

[11] T. A. Arias, M. C. Payne, and J. D. Joannopoulos, Phys. Rev. Lett. **69**, 1077 (1992).

[12] G. W. Fernando, G.-X. Qian, M. Weinert, and J. W. Davenport, Phys. Rev. B **40**, 7985 (1989); N. Chetty, M. Weinert, T. S. Rahman, and J. W. Davenport, Phys. Rev. B **52**, 6313 (1995).

[13] J. M. Holender, M. J. Gillan, M. C. Payne, and A. D. Simpson, Phys. Rev. B **52**, 967 (1995).

[14] N. Marzari, Ph.D. Thesis, University of Cambridge, 1996; N. Marzari *et al.* (to be published).

[15] M. J. Gillan (private communication).

[16] The mixing factor used to conjugate a direction depends on the scalar product between the original and the preconditioned gradient; we instead use the latter on both sides, without noticing a deterioration of convergence.

[17] The cost of one ensemble-DFT iteration (two f_{ij} iterations per inner loop) ranges between $2/3$ and 2, in ABV units; the extrema are reached in the Fourier-limited regime, with or without the $\{\psi_i\}$ stored in real space. In the cubic-scaling regime, the cost is $\sim 3/2$.